Unknown Input Observers Breaking Confidentiality of Controller States

Enno Breukelman and Henrik Sandberg

I. Motivation

The ubiquitous implementation of Cyber-Physical Systems (CPSs) and the potential to remotely cause physical damage makes them attractive to cyber-attacks. In those attacks, adversaries primarily target the communication networks, intercepting the communication between the plant and the controller.

Traditionally, cyber-attacks are classified by three dimensions: confidentiality (concealment of information), integrity (trustworthiness of data), and availability (ability to use information or resources). More recently, [2] defined an attack space for cyber-physical systems, spanned by model knowledge, disclosure, and disruption resources. In this paper, we focus on *disclosure attacks*, in which an adversary gathers sensitive information from the controller, thus breaking its confidentiality. Disclosure attacks can be part of a larger attack scheme, where the attacker initially remains hidden until it uses its disruptive resources. These disruptions, such as denial of service or false data injection attacks, then compromise the integrity and availability of the information in the CPS.

Stealthy attacks pose a great threat to CPSs, where stealthiness refers to the attacker not triggering any alarm while possibly causing damage. To stay undetected, the attacker must reconstruct or manipulate the state in the anomaly detector, which in turn requires knowledge of internal controller states.

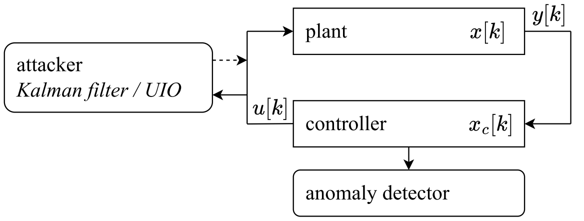

Figure 1: An attacker performs a confidentiality attack by reading control inputs to the plant $u[k]$. It uses a Kalman filter or a UIO to estimate the controller states $x_c[k]$. An anomaly detector raises an alarm if such an attack is detected.

In a control loop such as the one shown in Figure 1, there are two links in the communication network that an attacker can target: plant measurements $y[k]$ for *sensor attacks* and control inputs $u[k]$ for *actuator attacks*. In this work, we consider the case of an actuator attack, which poses an extraordinary threat to the CPS due to its direct connection to the plant.

II. Problem formulation

We consider a linear discrete-time plant and controller setup, where the two parts are interconnected in feedback by the plant measurements $y[k]$, controller inputs $u[k]$, plant states $x[k]$, and controller states $x_c[k]$. We further assume that the closed-loop system is internally stable.

$$\boxed{\begin{aligned} x[k+1] &= Ax[k] + Bu[k] + w[k] \\ y[k] &= Cx[k] + v[k] \end{aligned}}$$

$$\uparrow u[k] \qquad\qquad\qquad\qquad y[k] \downarrow$$

$$\boxed{\begin{aligned} x_c[k+1] &= A_c x_c[k] + B_c y[k] \\ u[k] &= C_c x_c[k] + D_c y[k] \end{aligned}}$$

We consider the problem where an adversary tries to gain access to the controller states using the control inputs $u[k]$. We write the problem description as follows:

Estimation problem

Estimate $x_c[k]$ by $\hat{x}_c[k]$ perfectly:

1. without bias: $\mathbb{E}[x_c[k] - \hat{x}_c[k]] = \mathbb{E}[e[k]] = 0$.
2. with zero steady-state error covariance:

$$\lim_{k\to\infty} \mathbb{E}\Big[e[k]\ e[k]^\top\Big] = \lim_{k\to\infty} \Sigma_c[k] = 0.$$

The attacker has an initial covariance $\Sigma_c[0] \succ 0$ and access to *controller inputs* to the plant $u[k]$.

In this work, we use two attack strategies to solve the estimation problem. For the first attack strategy, we adapt the procedure from [3], in which a sensor attack was performed using a Kalman filter. The second attack strategy is utilizing tools from delayed system inversion, i.e., using an Unkown Input Observer (UIO) to perform a delayed estimate of the controller states.

III. Strategy #1: Kalman Filter

To be able to apply the Kalman filter, we first combine the plant and controller states $z[k] = \Big[x[k]^\top x_c[k]^\top\Big]^\top$, and then construct the closed-loop state-space system with

$$z[k+1] = A_z z[k] + \alpha[k], \text{ where} \tag{1}$$

$$A_z = \begin{bmatrix} A + BD_cC & BC_c \\ B_cC & A_c \end{bmatrix},\ \alpha[k] = \begin{bmatrix} w[k] + BD_cv[k] \\ B_cv[k] \end{bmatrix}.$$

We can also write $u[k] = C_z z[k] + \beta[k]$ with the closed-loop output matrix $C_z = [D_cC\ \ C_c]$ and noise $\beta[k] = D_cv[k]$. Due to the random noise, the closed-loop state is a random variable. The conditional probability distribution of the closed-loop states $z[k+1]$ given a sequence of control inputs $\{u[i]\}_{i=0}^k$ under the influence of random noise, reads as

$$z\Big[k+1 \mid \{u[i]\}_{i=0}^k\Big] \sim \mathcal{N}(\hat{z}[k+1], \Sigma_z[k+1]).$$

The Kalman filter is an unbiased estimator, which means that $\mathbb{E}[z[k]] = \hat{z}[k]$. Therefore, to estimate the controller states per-

An extended version of this abstract has been presented at the Conference for Decision and Control 2024 in Milano, Italy [1].

fectly, we need to show that the steady-state covariance Σ_∞ is of the following form

$$\lim_{k\to\infty} \Sigma_z[k] = \Sigma_\infty = \begin{bmatrix} P & 0 \\ 0 & 0 \end{bmatrix}, \text{ with } P \succcurlyeq 0. \tag{2}$$

A steady-state covariance of this form implies zero covariance for the estimate of the controller states, $\lim_{k\to\infty} \Sigma_c[k] = 0$.

To be able to obtain an error covariance of the shape (2), we need to show two things: (a) Σ_∞ is the *unique* and *strong* solution of the Kalman filter's Riccati equation, and (b) we obtain *exponential convergence* of Σ_z towards Σ_∞.

Our first main result then reads as

Theorem 1 (Exponential convergence of Kalman filter): Suppose that the attacker has access to all model parameters of the plant and the controller. Suppose further, that D_c has full row rank that $R = D_c \Sigma_v D_c^\top$ is invertible, $(A, D_c C)$ is detectable, and $z[0]$ is uncorrelated with $\alpha[k], \beta[k]$.
Then, the attacker's controller estimate converges exponentially, if and only if $\rho\big(A_c - B_c D_c^\dagger C_c\big) < 1$ and $\Big(A, \Sigma_w^{1/2}\Big)$ does not have unreachable modes on the unit circle.

Note, that the attacker has access to control inputs $\{u[i]\}_{i=0}^{k}$ to estimate the controller states at time step $x_c[k+1]$.

IV. Strategy #2: Unknown Input Observer

The control inputs to the plant over a range of $L+1$ time steps, denoted as $[k : k+L]$, can be written as

$$u[k : k+L] = \mathcal{O}_L x_c[k] + \mathcal{J}_L y[k : k+L], \tag{3}$$

with the observability matrix $\mathcal{O}_L$ and the invertibility matrix $\mathcal{J}_L$. These matrices are computed recursively as

$$\mathcal{O}_L = \begin{pmatrix} C_c \\ \mathcal{O}_{L-1} A_c \end{pmatrix}, \quad \mathcal{J}_L = \begin{pmatrix} D_c & 0 \\ \mathcal{O}_{L-1} B_c & \mathcal{J}_{L-1} \end{pmatrix},$$

with $\mathcal{O}_0 = C_c$, and $\mathcal{J}_0 = D_c$. The matrices are computed together with the system inherent delay L. An observer for the controller states that operates independently of plant measurements reads as

$$\hat{x}_c[k+1] = E\hat{x}_c[k] + Fu[k : k+L]. \tag{4}$$

To motivate the design of E and F, we investigate the estimation error $e_c[k] = x_c[k] - \hat{x}_c[k]$, using (3) and (4):

$$\begin{aligned} e_c[k+1] = Ee_c[k] &+ (E - A_c + F\mathcal{O}_L)x_c[k] \\ &+ F\mathcal{J}_L y[k : k+L] - B_c y[k]. \end{aligned} \tag{5}$$

For the error to converge to zero, matrix E needs to be stable and F needs to fulfill both of the conditions

$$F\mathcal{J}_L = [B_c \; 0 \; \dots \; 0] \;\text{ and }\; E = A_c - F\mathcal{O}_L. \tag{6}$$

The first part in (6) ensures the independence of the error from the unknown plant measurements, while the second part ensures independence from the controller states. Theorem 3.2 in [4] shows that an UIO exists if and only if the controller is strongly detectable, i.e., it holds that

$$\operatorname{rank} \begin{bmatrix} A_c - zI & B_c \\ C_c & D_c \end{bmatrix} = n_c + n_y, \forall z \in \mathbb{C}, |z| \geq 1. \tag{7}$$

This condition coincides with all invariant zeros of the controller having a magnitude of less than one.

Theorem 2 (Exponential convergence of UIO): Suppose that the attacker has access to the model parameters of **only** the plant.
Then, the attacker's controller state estimate converges exponentially fast, if and only if it is minimum-phase (7).

Here, the attacker has access to control inputs $\{u[i]\}_{i=0}^{k+L}$ to estimate the controller states at time step $x_c[k+1]$.

V. Summary

The Kalman filter requires information on the plant, controller, and noise statistics, while the UIO only requires information on the controller. Additionally, the Kalman filter requires the closed-loop system to be stable. In contrast, the UIO estimates the controller's states directly and is therefore independent of the closed-loop dynamics.

The Kalman filter is designed to compute a one-step ahead estimate, based on the current control input. If $D_c = 0$, the noisy measurement signal $y[k]$ does not instantaneously act on the control input $u[k]$, which is what the Kalman filter bases its prediction on. Only after a delay of one time step $L_{\min} = 1$ does the noise act on the controller states and, therefore, on the control input $u[k]$. Thus, the Kalman filter fails to predict the controller states perfectly if the noise does not directly act on the control inputs. Requiring D_c to be of full row rank for the Kalman filter to converge implies that $n_y \geq n_u$. This condition is not restrictive since a subset of control inputs can be picked if violated. Contrarily, the UIO needs to have at least as many measured outputs as unknown inputs $n_u \geq n_y$.

In [3], it is shown that an attacker using a Kalman filter and plant measurements $y[k]$, require the controller to not have unstable poles. In contrast, we find that if an attacker accesses the control inputs $u[k]$, Kalman filter and UIO require stable controller zeros.

More specifically, regarding the Kalman filter, we obtain from Theorem 1 that the instantaneous right-inverse exists and is stable. For the UIO, Theorem 2 is more general and requires all invariant zeros to be stable. This implies, that the L-delay left-inverse of the controller exists and only has stable poles.

References

[1] E. Breukelman and H. Sandberg, "Unknown Input Observers Breaking Confidentiality of Controller States," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, Milan, Italy: IEEE, Dec. 2024, pp. 2373–2378. doi: 10.1109/CDC56724.2024.10886707.

[2] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*, Beijing China: ACM, Apr. 2012, pp. 55–64. doi: 10.1145/2185505.2185515.

[3] D. Umsonst and H. Sandberg, "On the confidentiality of controller states under sensor attacks," *Automatica*, vol. 123, p. 109329–109330, 2021, doi: https://doi.org/10.1016/j.automatica.2020.109329.

[4] S. Sundaram, "Fault-Tolerant and Secure Control Systems," *Lecture Notes, Department of Electrical and Computer Engineering, University of Waterloo.*